DOCUMENT RESUME

ED 416 631 EC 306 182

AUTHOR	Thurlow, Martha; Hurley, Christine; Spicuzza, Richard; El Sawaf, Hamdy
TITLE	A Review of the Literature on Testing Accommodations for Students with Disabilities. State Assessment Series: Minnesota Report 9.
INSTITUTION	National Center on Educational Outcomes, Minneapolis, MN.; Minnesota State Dept. of Children, Families, and Learning, St. Paul.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	1996-08-00
NOTE	43p.
CONTRACT	R279A50011
AVAILABLE FROM	Publications Office, NCEO, 350 Elliott Hall, 75 East River Road, University of Minnesota, Minneapolis, MN 55455; phone: 612-626-1530; fax: 612-624-0879; World Wide Web: http://www.coled.umn.edu/nceo (document may be copied without charge, additional print copies $10).
PUB TYPE	Information Analyses (070)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Disabilities; *Educational Assessment; Educational Policy; Educational Trends; Elementary Secondary Education; Federal Legislation; Federal Programs; Legal Problems; *Psychometrics; Research Needs; Statistical Analysis; Student Attitudes; *Student Evaluation; Teacher Attitudes; Test Validity; *Testing Problems; Trend Analysis
IDENTIFIERS	*Minnesota; *Testing Accommodations (Disabilities)

ABSTRACT

This report reviews the current literature (mostly since 1993) on testing accommodations for students with disabilities, with an emphasis on studies examining the effects of testing accommodations on the technical integrity of assessment measures. The report notes a continuing lack of empirical research on testing accommodations, but evidence of changes in policy (especially enactment of the Americans with Disabilities Act and implementation of the National Education Goals) suggests that accommodations will receive more direct attention. The recent federal funding of projects to examine issues related to assessment for students with disabilities is also noted, as is the increasing number of journal articles, books, and professional documents about testing and accommodations. The report is organized into five sections: (1) a brief description of the methodology used to conduct the literature review; (2) empirical studies of testing accommodations; (3) legal considerations related to testing accommodations; (4) teacher and student perceptions of testing accommodations and modifications; and (5) conceptual issues. Appended are a review of early studies in testing accommodations conducted by the American College Testing Program and the Educational Testing Service and a listing of relevant research projects currently supported by the U.S. Department of Education. (Contains 39 references.) (DB)

****

ED 416 631

STATE ASSESSMENT SERIES
Minnesota Report 9

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A Review of the Literature on Testing Accommodations for Students with Disabilities

BC 30682

Minnesota Children

Minnesota Department of Children, Families & Learning

STATE ASSESSMENT SERIES
Minnesota Report 9

A Review of the Literature on Testing Accommodations for Students with Disabilities

Minnesota Assessment Project

Project Staff:
Constance Anderson • Cathy Wagner • Michael Tillman
Minnesota Department of Children, Families and Learning

Project Advisors:
Karon Hergenhahn • Bounlieng Phommasouvanh • Marilyn Sorensen • Elizabeth Watkins
Minnesota Department of Children, Families and Learning

Prepared By:
Martha Thurlow • Christine Hurley • Richard Spicuzza • Hamdy El Sawaf
University of Minnesota

August 1996

The Minnesota Assessment Project is a four-year, federally funded effort awarded to the Minnesota Department of Children, Families and Learning from the U.S. Department of Education, Office of Educational Research and Improvement. The project's goal is to promote and evaluate the participation of students with limited English proficiency and students with disabilities in Minnesota's Graduation Standards. Specifically, the project will examine ways in which students with limited English and students with disabilities can participate in the Basic Standards Exams of reading, mathematics and written composition and in the performance-based assessments of the high standards in the Profile of Learning.

This project is supported, in part, by a grant to the Minnesota Department of Children, Families and Learning from the U.S. Department of Education, Office of Educational Research and Improvement (Grant #R279A50011). Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

Permission is granted to copy this document without charge. Additional print copies may be purchased for $10.00 from:

NATIONAL CENTER ON EDUCATIONAL OUTCOMES
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
612.624.8561 • Fax 612.624.0879
http://www.coled.umn.edu/NCEO

This document is available in alternative formats upon request.

Overview

In 1993, the National Center on Educational Outcomes (NCEO) published a comprehensive literature review on testing accommodations for students with disabilities. In it, Thurlow, Ysseldyke, and Silverstein reviewed empirical studies, namely those conducted by the Educational Testing Service (ETS) and by the American College Testing (ACT) Program (see Appendix A for a summary of these early studies). They also addressed policy and legal considerations, technical concerns, minimum competency and certification/licensure testing efforts, and existing standards and accommodations allowed in state assessment systems. The results of their review documented that very little empirical research existed on testing accommodations and revealed that there was tremendous variability across states in terms of the degree to which they included students with disabilities in assessment or made accommodations for them. A review of the literature published since that report suggests that in some ways, little has changed with respect to empirical research on testing accommodations. Currently, comprehensive empirical studies of the effects of testing accommodations are still noticeably absent from the literature on assessment and students with disabilities.

There are, however, several indications of change. Much of this change comes in the form of policy: the enactment of the Americans with Disabilities Act (ADA; PL 101-336), the implementation of the National Education Goals, and the increasing use of high stakes assessments in many states. These policies assure that the issue of how to make appropriate accommodations for students with disabilities will receive more direct attention. Also, federal funds have begun to be directed toward these efforts. In 1995, the U.S. Office of Special Education Programs funded three projects to examine issues related to assessment for students with disabilities. Similarly, the U.S. Department of Education's Office of Educational Research and Improvement (OERI) funded eight states, including Minnesota, and one consortium of 22 states, to improve their state assessments through alignment of assessments with standards, and increased inclusion of students with disabilities and limited English proficiency in their assessments (Erickson, Thurlow, & Ysseldyke, 1996). Most of these projects address testing accommodations. Appendix B provides a list of project titles and recipient organizations. Additional projects are to be funded in the future. Another indicator of change is the ever-increasing number of journal articles, books, and professional documents that have been written about testing and accommodations. Although we still have very little data, one result of this increasing literature base is that we are deepening our understanding of the critical issues, which in turn should help inform the emergent research. The purpose of this report is to provide an updated review of the literature on testing accommodations for students with disabilities, with a particular emphasis on studies examining the effects of testing accommodations on the technical integrity of assessment measures. Like the original NCEO

report, the goal is to answer the question, "What do we currently know about testing accommodations for students with disabilities?"

This report is organized into five sections: (1) a brief description of the methodology used to conduct the literature review, (2) empirical studies of testing accommodations, (3) legal considerations related to testing accommodations, (4) teacher and student perceptions of testing accommodations and modifications, and (5) conceptual issues related to testing and accommodations.

Methodology

Sources of information for this literature review were wide-ranging and included books, journal articles, agency reports, personal communications with researchers involved in similar efforts in the field, documents published by research centers (e.g., North Central Regional Educational Laboratory [NCREL]) and testing companies (e.g., Educational Testing Service [ETS]), as well as papers presented at national conferences. During initial searches, the criteria for inclusion were (1) publication during or after 1993, and (2) if published prior to 1993, the source had not been included in Thurlow et al. (1993). Searches were conducted on educational and psychological computer databases – ERIC and PsycLit, the World Wide Web (using Alta Vista and Yahoo search engines), as well as a search of the Outcomes-Related Bank of Informational Text (ORBIT), a computerized literature database maintained by NCEO. The following keywords, listed here in alphabetical order, were used in various combinations to conduct database searches: accommodations, adaptations, assessment, competency tests, disabilities, effectiveness, empirical studies, graduation standards, high stakes assessment, measurement, modifications, psychometric properties/qualities, reliability, special education, standards, technical adequacy, test(s, ing), and validity. Finally, the Social Science Citation Index was reviewed to determine which authors have cited a previously published report or study.

With respect to language and terminology, people-first language is employed throughout this report regardless of the label used in the original document. However, in recognition of the ongoing debates surrounding definitions of accommodations, adaptations, and modifications, we use the term chosen by the original author.

Empirical Studies of Testing Accommodations

The current review of the literature reveals only six studies that examined the effects of various accommodations or modifications; they are a heterogeneous group. Sample sizes ranged from very small (N=3) to quite large (N = over 17,000); subjects ranged in age from fourth graders to post-secondary aged; and the purposes of the studies were quite varied. Three of the studies examined the effects of timing accommodations on student performance, two explored the effects of different format modifications on tests, and the sixth study examined the effects of modifications to curricular activities on problem and on-task behaviors. Each of the studies is reviewed below.

Timing Accommodations

Prior studies of special test administrations (i.e., nonstandard administrations that may include

changes in presentation, response, or testing environment) of the Scholastic Aptitude Test (SAT) conducted by the College Board and the Educational Testing Service (ETS) (Willingham et al., 1988) showed that overall, special and regular administrations of the SAT are comparable (in terms of reliability, construct validity, and predictive validity) with the exception of timing accommodations. One of the major conclusions from this body of research was that attempts should be made to establish empirically-based timing conditions for special administrations.

Funded by the College Board and ETS to address this issue, Ragosta and Wendler (1992) sought to establish empirically derived testing times for special administrations of the SAT for students with disabilities and to develop eligibility guidelines for special test administrations. Using data from the 1986-87 and 1987-88 SAT test administration timing records, the SAT history files, and a survey questionnaire, their sample included over 17,000 students with disabilities who took special administrations of the test. Students with learning disabilities accounted for nearly 80% of the sample, followed by students with visual impairments (9%), hearing impairments (4%), physical disabilities (4%), and students with multiple disabilities (2%). The researchers focused on "Plan A" special administrations in which students are allowed to take the test over two consecutive testing days for no more than six hours of testing time per day. Standard administrations are given in one day for a total of 2.5 hours.

In terms of comparable time limits, the researchers found that in general, comparable numbers of students with and without disabilities completed the exam when students with disabilities were given one and a half to two times the standard testing time. Two- to three-times the standard time was needed for students with visual impairments who required Braille or cassette tape administrations; students who were deaf or hard-of-hearing required somewhat less than double time. A related question was whether there were other groups of students who might require extraordinary testing time. Results indicated that beyond the above mentioned students, the only others who were likely to need more than double time were students with multiple disabilities.

The researchers also examined testing time by sections of the test and found that students with hearing impairments took the least amount of time to complete a section, students with visual impairments using Braille or cassette versions of the test required the most time, and all the other disability groups fell somewhere in between. In addition, they found that there was an uneven distribution of time across sections of the test; students with disabilities spent more time on the first section of the test than on any other.

The second purpose of the Ragosta and Wendler (1992) study was to establish more stringent eligibility criteria for special test administrations. SAT eligibility criteria include having a current IEP or two signed documents describing the nature of the disability, how it was

diagnosed, and stating that the disability meets state guidelines for certification (i.e., eligibility for special education). Ragosta and Wendler developed a hierarchy of eligibility based on school practices that ranged from "certain" to qualify to "doubtful." For example, students with current IEPs or who attend special schools or classes for students with disabilities would qualify for accommodations, whereas qualification would be doubtful for students with no documentation of needing accommodations in school. Using this hierarchy, the researchers found that they could differentiate the severity of disability for students with hearing and visual impairments and the type of disability for students with physical disabilities. However, no distinctions could be made for students with learning disabilities.

Ragosta and Wendler (1992) also noted two consistent problems when they examined testing records for the students who took special administrations: (1) there was a significant proportion of students with disabilities who certainly would have qualified for accommodations but took the standard administration, and (2) there was another significant number of students for whom it was doubtful that they would qualify but they had taken a special administration of the test. With respect to the first group, the authors stated that while it was possible that each of the qualifying students knew their options and elected to take the standard test, there is sufficient evidence to suggest that students with disabilities are sometimes uninformed about special test accommodations (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Evaluation [NCME], 1985; Ragosta, 1980). Regarding those students who did not qualify based on school practices, the authors contended that the current criterion of requiring outside documentation is inequitable, since families with greater economic resources are better able to secure these types of documents than are poorer families. Thus, eligibility criteria based on actual school practices appear to be needed.

In summary, the results of Ragosta and Wendler's (1992) study indicate that one-and-a-half to two-times the standard testing time would result in comparable numbers of students with and without disabilities completing the SAT, for most students with disabilities. For students using Braille or cassette versions of the test, or for students with multiple disabilities, two- to three-times the standard time is needed. Section timing, even for special administrations, appears necessary in order to make scores more comparable. With respect to eligibility guidelines, a hierarchy based on school practices was successful in differentiating between some types and levels of disabilities but was least successful for students with learning disabilities. The authors assert that eligibility criteria based on school practices would be more equitable.

The other two investigations of timing accommodations are not quite as rigorous or as detailed as Ragosta and Wendler's study. Both Munger and Loyd (1991) and Perlman, Borger, Collins, Elenbogen, and Wood (1996) used the Iowa Tests of Basic Skills (ITBS) to measure student performance under timed and non-timed conditions.

In the Perlman et al. study, 28 fourth grade and 57 eighth-grade students with learning disabilities were given the ITBS in either a timed (40 minutes) or untimed (2.5 hours) administration. Analysis of covariance results, using previous ITBS scores as the covariate, indicated that there were significant main effects for both timing and grade; students in the untimed condition and students in eighth-grade scored significantly higher. Additional findings suggested that the post-test was more reliable when untimed; students in the untimed condition did not always use all of the allotted time, and older students were more likely to need extra time. Moreover, fourth graders in the untimed condition tended to score higher than the fourth graders in the timed condition, even though they both used about the same amount of time. This result caused the authors to speculate as to whether the critical variable is time, or perhaps the reduced stress and more positive expectations that accompany students' knowing that they have unlimited time. The authors also suggested that providing empirically derived standards may be more comparable than just providing unlimited time. Unfortunately, the validity of findings is tempered by methodological concerns (e.g., non-random assignment of students to treatment conditions) that may limit generalization.

In the third study, 220 fifth-grade students with and without disabilities were administered the Language Usage and Expression and Mathematics Concepts subtests of the ITBS (Munger & Loyd, 1991). Students with disabilities included those with learning disabilities and students with physical disabilities (e.g., neurological or orthopedic) who were capable of taking the test independently under timed conditions. Each student took two forms of either the Language or Math subtest; one was timed (standard time), the other untimed (students were given as much time as needed to complete the test). The order of the test forms was varied and the procedure for testing was such that students took the first test, then a short break, and then took the second test. All students used large format answer sheets. Data analysis consisted of a two-group discriminant analysis and two-factor mixed analysis of variance (ANOVA). Results indicated that for both the Language and Math tests, students with disabilities could not be distinguished from students without disabilities based on completion or noncompletion of 90% of the test items or number of items attempted. In addition, there were no significant differences between groups when timing conditions were varied. The authors concluded that timing appeared to have had "little effect on the performance of either group" (p. 57). Based on their results, they suggested that many students with physical or learning disabilities need not be exempted or excluded from standardized testing.

While this study is interesting because of the younger age group, and because timing had little to no effect, there are significant limitations in the study design, making generalizations about the results tenuous. The most important limitation was that only students who were capable of taking the tests independently under timed conditions were included.

A summary of the major findings from each of the three studies is provided in Table 1. There are several points to be highlighted from the research on timing accommodations. Empirically-derived standards are recommended over unlimited time. Furthermore, Ragosta and Wendler's study suggests that there are ways to determine comparable time limits. Given the varying effects of timing modifications in the three studies, more research is needed, particularly methodologically sound investigations that test the effects of timing accommodations in ways that are generalizable.

Format Modifications

Format modifications involve two general categories reflecting either changes in the presentation of materials (e.g., Braille or audiocassette editions, large-print tests) or in the mode of response (e.g., give response in sign language, mark responses in a test booklet) (NCEO, 1993). In the two studies presented here, one examined changes in presentation format (Dalton, Morocco, Tivnan, & Rawson, 1994) while the other investigated the effects of both presentation and response format changes (Mick, 1989).

Dalton et al. (1994) examined the effects of two alternative assessments — a constructed diagram test and a written questionnaire — on 172 fourth-grade students with (N=33) and without (N=139) learning disabilities from six urban and two suburban classrooms. All of the students had participated in a hands-on science curriculum. Results indicated that students' outcomes were a function of learner status (learning disability, low achieving, average achieving, and high achieving) and level of science knowledge after instruction. More relevant for this review was the finding that students with learning disabilities, and low and average achieving students, obtained higher scores on the constructed diagram test than on the questionnaire after controlling for domain-specific knowledge. High achieving students performed comparably on the two measures. The majority of students (88%) reported that they liked the diagram test better, stating that it was fun and easier than the questionnaire. Possible explanations for differential performance included the hypothesis that the two tests measured different aspects of achievement or that the diagram test scaffolds student performance by focusing attention, activating relevant schema, and providing a more constrained response.

Mick (1989) examined the effects of three format modifications to the Instructional Objectives Exchange (IOX) Basic Skill Test (reading subtest, secondary level) on the achievement of 76 secondary students with learning disabilities and mild-to-moderate mental handicaps. The modifications were (a) moderately increased print size, (b) use of unjustified lines for right margins, and (c) responses recorded on test booklets rather than answer sheets. Using a repeated replication design, the unmodified and modified versions of the test were administered

Table 1: Findings from Studies of Timing Accommodations

Study	Sample	Finding
Ragosta and Wendler (1992)	17,000 students with disabilities who took special administrations of the SAT	1) Comparable numbers of students with and without disabilities completed the exam when students with disabilities were given one and one-half to two times the standard testing time. 2) Two to three times the standard time was needed for students using Braille or cassette versions of the test, or for students with multiple disabilities.
Perlman, Borger, Collins, Elenbogen, and Wood (1996)	28 fourth grade and 57 eighth grade students with learning disabilities who took the ITBS under timed and untimed conditions	1) Students in untimed condition scored significantly higher than students in timed condition. 2) Students in the untimed condition did not always use all of the allotted time. 3) Older students were more likely to need extra time. 4) Fourth-graders in untimed condition scored higher than fourth-graders in timed condition.
Munger & Loyd (1991)	220 fifth-grade students with and without disabilities who took the Language Usage and Expression and Mathematics Concepts subtests of the ITBS under timed and untimed conditions	Timing had little to no effect on the performance of students with or without disabilities.

to each student. Results indicated that both students with learning disabilities and mild-moderate mental disabilities performed significantly better on the unmodified version of the test. One explanation put forth by the author was that secondary students become test-wise after long-term exposure to standardized test formats, including answer sheets and justified margins. However, when one considers the numbers of students who actually passed either version of the test (with a criterion of 70%), the practical significance of Mick's findings are called into question. Less than half of the students with learning disabilities and only two students with mild-moderate mental handicaps passed either version of the test. Passing scores on both versions were obtained by only eight students with learning disabilities and none of the students with mild-moderate mental disabilities.

It is hard to draw conclusions about format modifications based on these two studies. It appears that format changes can affect student performance although the mechanism by which this happens is still unknown. Further research based on Dalton et al.'s work might explore whether similar results are found for older students, whether results change if students have not participated in a hands-on science curriculum, and what, if any, relationship exists between students' perceptions of exams as fun and their performance. Regrettably, neither study addressed validity issues related to making format modifications.

Modifications to Curricular Activities

Dunlap, Foster-Johnson, Clarke, Kern, and Childs (1995) sought to determine whether problem behaviors in three elementary aged students with disabilities (including autism, mental retardation, and emotional/behavioral disabilities) could be reduced and on-task behavior increased if students' curricular activities were modified according to their own interests. For each student, a particular instructional objective was held constant, but the way in which the objective was met was modified to make the task more interesting to the student. Information about students' interests were obtained from a variety of sources, including teacher input, classroom observations, directly asking students, and by conducting brief probes. Using a reversal design, results showed that all three students reduced problem behaviors and increased on-task behaviors when their curricular tasks were modified according to their interests. In their discussion, the authors assert that although the conceptual basis for the changes in student behavior is not fully understood at this time, the functional outcomes are what is important.

Although not directly relevant to the issue of testing accommodations, particularly high stakes testing, this study was included because there appears to be a theme emerging from several of the other studies in this section, including Dunlap et al. (1994), indicating that student interest and level of comfort may be important variables related to performance. More research is needed to determine whether, in fact, these are relevant variables.

Summary

In this section we reviewed six studies examining the effects of timing, format, and curricular accommodations or modifications. Because the studies were divergent in both their purposes and methodologies, so too, are the results and conclusions to be drawn from them. Overall, the findings from this group of studies make absolutely clear the need for rigorous empirical research. While some of the studies were technically sound, others were not and the end result is that we still have very little data on the effects of testing accommodations. In the absence of a comprehensive empirical base from which to make decisions about testing accommodations, other related literature bases must be consulted. In the next section, legal issues related to testing accommodations are described.

Legal Considerations Related to Testing Accommodations

The primary source of information on legal issues pertaining to testing accommodations is Dr. S.E. Phillips, a professor at Michigan State University. Specializing in legal issues in assessment and psychometrics, Dr. Phillips also holds a law degree. She is the author of all four of the articles reviewed in this section. Understandably, there is considerable overlap across the articles. In each article Phillips reviews (in various levels of detail) the federal statutes and case laws related to the topic of educational testing and accommodations. She also discusses psychometric considerations related to testing accommodations. Phillips (1996) notes that although new legal challenges regarding testing accommodations are most likely to emerge from the ADA legislation enacted in 1992, as yet, there has been no definitive case law. Therefore, the reader is referred to Thurlow et al. (1993) for descriptions and reviews of the relevant federal statutes and cases. This paper will not review these decisions and laws except where necessary to illustrate a point. The purpose of this section is to review the relevant issues related to making testing accommodations, describe where legal challenges are likely to be made, and provide recommendations for decision making.

Legal Issues

According to Phillips (1993, 1996), the core issue with respect to testing accommodations is one of balance: balancing the rights of the individual student with a disability with the need to maintain the validity of the assessment tool used to measure student performance. While the important issue for educators, students, and parents is most likely the protection of the student and the opportunity to participate to the best of his or her ability, "the bottom line for measurement specialists is validity – are scores with and without accommodations comparable" (Phillips, 1994, p. 96).

Federal statutes and requirements for practice set forth by professional organizations provide guidance as to what is required and what is merely expected in terms of testing accommodations for students with disabilities. The ADA mandates that private entities uphold Section 504 of the Rehabilitation Act (1973), which requires that "no otherwise qualified handicapped individual shall, solely by reason of his handicap, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving Federal financial assistance" (cited in Phillips, 1994, p. 106). However, the U.S. Supreme Court, in Southeastern Community College v. Davis, held that Section 504 does not require "an educational institution to lower or to effect substantial modifications of standards to accommodate a handicapped person" (cited in Phillips, 1993, p. 372). Additional guidance comes from the Standards for Educational and Psychological Testing (AERA, APA, NCME, 1985) which state that

> unless it has been demonstrated that the psychometric properties of a test...are not altered significantly by some modification, the claims made for the test...cannot be generalized to the modified version... When tests are administered to people with handicapping conditions, particularly those handicaps that affect cognitive functioning, a relevant question is whether the modified test measures the same constructs. Do changes in the medium of expression affect cognitive functioning and the meaning of responses? (cited in Phillips, 1993, p. 381)

This issue is particularly challenging for students who have "mental disabilities" (e.g., learning disabilities and other cognitive disorders) (Phillips, 1994). Accommodations for students with physical disabilities (e.g., blindness, confinement to a wheelchair) usually involve the removal of physical barriers or changes in presentation format that do not affect the validity of the test (e.g., Braille versions), and issues around outside verification or documentation of the disability are generally non-existent. That is not the case for learning and other cognitive disabilities. It has been shown that the classification of learning disabilities is a somewhat arbitrary process, dependent largely upon "the method used to identify the disability, the availability of services in particular disability categories, and the perception of the parent(s) of the benefit of special education for that student" (Phillips, 1994, p. 114). Furthermore, other previously described research (Ragosta & Wendler, 1992) indicated that beyond initial classifications, differentiation of students with learning disabilities, by subgroup (reading or math disability) or by level of severity, is quite difficult.

One result of these challenges is that measurement issues are complicated when testing accommodations for learning disabilities are considered, particularly since some of the accommodations are likely to affect the meaning and interpretation of scores (Phillips, 1994). For students with cognitive disabilities, Phillips states, "because the disability is often

intertwined with the skills the test user wishes to measure, allowing the accommodation may effectively exempt the disabled person from demonstrating the mental skills the test measures" (p. 95).

Legal Challenges and Criteria for Standards Assessments

Phillips (1996) asserts that legal challenges from students denied testing accommodations are most likely to emerge as unlawful discrimination cases under the ADA. Despite the fact that this newer legislation is as yet untested in the federal courts and prior court cases have generally not drawn the line between valid and invalid accommodations (Phillips, 1994, 1996), what is known is that "in the past, judges have been deferential to academic decisions as long as proper procedural safeguards are followed. The courts have reinforced the quality issue; schools do not have to lower standards" (Phillips, 1995, p. 5).

In an article focusing on the legal defensibility of standards, Phillips (1996) describes legal criteria, drawn from prior assessment cases and statutory law, that may be challenged as states develop and implement high stakes assessments and graduation standards. Specifically, she defines and discusses six legal criteria for "descriptive standards" (i.e., goal statements describing what students should know and be able to do in specific content areas): notice, curricular validity, adverse impact, opportunity for success or fundamental fairness, articulating defensible standards, and assessment accommodations for students with disabilities. Each of these is reviewed below.

Notice. Notice is a procedural due process requirement. In general, the courts have asserted that students and their families must be provided adequate notice of any required assessment that may influence the receipt of a high school diploma (Debra P. v. Turlington, 1979-1984). Phillips states that while "adequate" has been defined differently in different court cases, the appropriate amount of time will likely be the amount of time needed to "provide students and school personnel with clear indications of the specific content (knowledge and skills) and performances for which they will be held accountable" (p. 6). Although she discusses the issue of time to learn the skills as part of the next criteria — curricular validity — she recommends that anytime changes are made to standards, "the notice period should probably be as long as that for the implementation of the original graduation standards" (p. 6). In her opinion, this will allow sufficient time for skills to be included in the curriculum and subsequently taught to students.

Validity. According to Phillips, "curricular validity requires assessment administrators to demonstrate that students have had an opportunity to learn the knowledge and skills included on a graduation assessment" (p. 6). Courts have held that relevant sources of evidence for establishing curricular validity include the inclusion of tested skills in the official curriculum

and that the majority of teachers assert that these are skills they should teach (Debra P.). It is recommended that districts collect these sources of information through multiple measures (e.g., student and teacher surveys, textbook and curricular guide reviews). In her discussion of curricular validity, Phillips notes several gray areas. For example, if complex multiple skills are included in a standard, it may be challenging to "identify the point in the curriculum where students are expected to have learned the content and/or performances necessary to demonstrate attainment of the standard" (p. 6). Other areas of concern include remediation of skills that can not be taught in a short time frame; cases where science or math standards involve writing but students have not been required to write in their math or science classes; and the prediction of performance on graduation standards based on performance on standards in prior grades.

Adverse impact. This requires that consideration be made to the potentially adverse impact of new graduation standards on historically disadvantaged groups. Phillips suggests that one argument that may arise with respect to performance-based assessment (if disadvantaged groups perform more poorly on such tests) is that the state or district has "discriminated by replacing a less discriminatory alternative (multiple choice test) with graduation standards that result in greater disadvantage" (p. 6). In the only related cases tried so far, courts have found that cost effective alternatives with less adverse impact must be considered in employment testing (Wards Cove Packing Co. v. Antonio, 1989).

Opportunities for success or fundamental fairness. Deriving from the substantive due process clause of the fourteenth amendment, fundamental fairness requires that "assessments must adhere to professional requirements, be valid, fair, avoid arbitrary or capricious procedures, and provide all students with conditions fostering an equal chance at success" (Phillips, 1996, p. 7). Professional recommendations from the Code of Fair Testing Practices (JCTP, 1988) and the Testing Standards (APA/AERA/NCME, 1985) also suggest equal-opportunities-for-success mandates. Phillips makes the important point that the due process requirement is "not a guarantee of equal outcomes but rather of standardized conditions which ensure that no student receives an unfair advantage or penalty" (p. 7), hence her coining of the term "opportunity for success." She provides several examples illustrating possible scenarios in which one group of students might have an unfair advantage or experience unfair penalties. These include the use of differential equipment, applying standards for individual students to group work, outside assistance, and procedural differences.

Articulation of defensible standards. Phillips contends that there are two major issues to consider with respect to this criterion: (1) standards should specify clearly observable behaviors, and (2) the issue of parental rights must be given thoughtful consideration. Phillips states that even when standards clearly address the first issue, "parents may demand the right to preview the content of multiple-choice questions or performance tasks administered to their children" (p. 8). Parental concerns addressed in prior court cases have been based on religious

convictions as well as concerns about students' rights to privacy or school pressure to support objectionable points of view (e.g., Maxwell v. Pasadena I.S.D., 1994). Issues raised in trying to balance parents' rights with test administration include assessment security, parents' constitutional rights, legislative action, and the conditions under which parents can review the test.

Assessment accommodations for students with disabilities. This criterion focuses on cognitive disabilities and addresses the issues of valid versus invalid accommodations (as defined by The Test Standards [AERA, APA, NCME, 1985] and relevant court cases), score notations for nonstandard accommodations, explicating assumptions, and accommodation alternatives. These will be discussed in greater detail in the following section on decision making for accommodation.

Thus far, we have examined what the issues are related to testing accommodations for students with disabilities, primarily those that are cognitive in nature. We have also explored what the relevant legal challenges are likely to be as graduation standards become the norm in many states and cases are tried under ADA. The question that remains, then, is what to do? In the absence of much empirical data on the effects of accommodations, what can administrators and test developers do to try to maintain the balance between the goals of full inclusion and maintaining test validity? This question is addressed in the next section on considerations and guidelines for decision making.

Considerations for Accommodations and Guidelines for Decision Making

As policy makers, administrators, and educators attempt to achieve a reasonable balance between individuals and tests, there are numerous issues to consider. While some requests for accommodations may be inappropriate, the courts have ruled that decisions about accommodations must be made on a case-by-case, individual basis (Hawaii State Department of Education, 1990). Phillips asserts that while generalizations about the accommodations for specific disabilities are not really possible, it is plausible to make some generalizations about the general appropriateness of "specific accommodations for a particular testing application" (Phillips, 1994, p. 98). In particular, she asserts that when determining whether a requested test accommodation is valid, the administrator or other decision maker should consider the purpose of the test, the skills to be measured, and the inferences to be made from the test score. To that end, she has developed a set of questions for people to reflect on when they are considering departing from standardized testing conditions (Phillips, 1993, 1994):

1. Will format changes or alterations in testing conditions change the skill being measured?

2. Will the scores of examinees tested under standard conditions have a different meaning from scores for examinees tested with the requested accommodation?

3. Would examinees without disabilities benefit if allowed the same accommodation?

4. Does the examinee with a disability have any capability for adapting to standard test administration conditions?

5. Is the disability evidence or testing accommodations policy based on procedures with doubtful validity or reliability?

A "yes" answer, according to Phillips, suggests that an accommodation is not appropriate. Another key consideration is what the objective of the standard requires. Phillips (1993) writes that "in judging the effects on content validity of deviations from standardized testing conditions, one must evaluate the intent of the objectives as they are currently written" (p. 382). Furthermore, decisions "regarding alternate passing scores and multiple formats should be made based on empirical data where feasible" (Phillips, 1996, p. 12). "In the long run, allowing all students access to useful accommodations may be fairer to low achieving students (p. 12).

In thinking about taking action, Phillips offers several possible options. One possibility is self-selection with informed disclosure (Phillips, 1994). In this scenario all reasonable accommodation requests (for any student) would be honored and then the accommodations, not the disability, would be noted on the student's transcript or diploma along with the passing grade. Provided that both parents and students are given advance notification of the notation procedure, are provided clear information about possible ramifications, and give permission, all students could theoretically access testing accommodations. Phillips asserts that this would relieve measurement specialists of the task of judging which disabilities qualify for accommodations as well as determining whether a student even qualifies for accommodations. The arguments against this practice are two-fold. First, the issue of flagging is controversial. On the one hand, many advocates for people with disabilities believe that notation or "flagging" unfairly labels a person as having a disability and denies them the opportunity to compete fairly with students without disabilities" (Phillips, 1993). Moreover, there is the potential for misuse of such notations such that a person with a disability might be wrongfully discriminated against (Phillips, 1994). Conversely, many test developers argue that "reporting scores from nonstandard test administrations without special identification violates professional principles, misleads test users, and perhaps even harms handicapped test takers whose scores do not accurately reflect their abilities" (AERA, APA, & NCME, 1985). Second, the cost of providing

accommodations to any student who requests them may be prohibitive both for large urban districts and small rural districts.

A second possibility is to eliminate extraneous skills so that accommodations would not be necessary (Phillips, 1994). She offers as examples making a "speeded" test nonspeeded for all examinees or designing a test to measure "communication" skills such that it could be administered in written or oral form. Obviously, the need to be certain that the eliminated skills are truly extraneous to the skill being measured is imperative in order to maintain the validity of test score interpretation.

No matter what decisions a program or district makes regarding the provision of testing accommodations, Phillips (1996) strongly asserts that "the most important requirement... is the development of a comprehensive written policy outlining the procedures for requesting accommodations and detailing how decisions will be made regarding specific requests" (1996, p.12). The list of guidelines for the development and implementation of legally defensible testing accommodation policies provided by Phillips (1996) is presented in Table 2.

Summary

In this section, legal issues related to testing accommodations were reviewed, including the challenges related to making accommodations for mental disabilities and the need to balance student inclusion with the need for valid measures. Legal challenges and criteria for standardized assessments were described and considerations for accommodations and guidelines for decision making were presented. In the following section, we review research focusing on the purveyors and consumers of accommodations — teachers and students.

Teacher and Student Perceptions of Testing Accommodations and Modifications

Another body of research related to the effects of testing accommodations examines the perceptions of those people who generally provide or receive classroom or testing modifications — teachers and students. According to Jayanthi, Bursuck, Havekost, Epstein, and Polloway (1994), "very often, the testing practices for individuals with disabilities in general education classes are a reflection of the testing policies established at the state and local school district levels" (p. 695). Understanding how teachers and students perceive and make decisions about testing and classroom accommodations is important for two reasons: (1) although the laws and guidelines are clear about the full inclusion of students with disabilities in both testing and instruction, there is not a lot of information about what the consumers (in this case, students) think about inclusion and modifications; and (2) if local practice is indeed a

Table 2: Recommendations for Legally Defensible Accommodation Policies (Phillips, 1996)

1	All districts, training programs, and applicants should be provided with written instructions for requesting accommodations.
2	A standardized form with a return deadline should be used to make accommodation requests.
3	Requesters must provide recent documentation of the disability by a licensed professional. Phillips included this criterion as a safeguard from the arbitrariness of the LD diagnosis. *Note:* Ragosta and Wendler's (1992) research suggests that the ability to use outside experts unfairly disadvantages people from lower SES backgrounds; therefore, another possibility is to use school-based criteria, such as a current IEP.
4	Related to number 3 above, requesters must provide documentation of any accommodations that have been provided in the requester's educational or training program.
5	If scores obtained under nonstandard conditions will be flagged or limited licenses granted, notify requesters of this fact, and ask them to sign a statement prior to testing confirming that they have been notified. If the examinee/requester is a minor, a parent or guardian should also sign.
6	A single individual within the testing agency should be designated to review and act on all requests for testing accommodations. A qualified consultant could provide assistance on borderline cases.
7	Testing accommodations requests should be reviewed on an individual, case-by-case basis, applying previously developed written criteria.
8	At the state or program level, collect data on accommodations for mental disabilities for which the effects on test validity are questionable.
9	Provide an expedited review procedure by the testing agency for all denied accommodation requests. Written decisions should be provided to the requester.
10	Upon written request, provide a formal appeal procedure, including a hearing, for requesters whose denials are upheld in the review process.
11	Under the IDEA, section 504, and the ADA, students probably cannot be asked to bear any of the additional costs of providing testing accommodations. Reasonable limitations of accommodations to specific testing dates and sites are probably acceptable.
12	Testing agencies may want to codify testing accommodations policies in administrative rules or legislation to ensure stability and consistency across changes in personnel.

reflection of broader belief systems, data collected from teachers may act as a barometer of larger societal views on including students with disabilities in educational settings.

Of the four studies reviewed here, three focused on teacher perceptions and one surveyed students. Teacher respondents were all general educators. The study of students did not indicate the nature of the sample (i.e., students with and without disabilities may have been included).

In a national study of 401 general education teachers, Jayanthi, Epstein, Polloway, and Bursuck (1996) examined teachers' perceptions of testing adaptations for students with disabilities in general education classrooms. Results indicated that for the majority of respondents (83%), general educators either alone or jointly with a special educator, were responsible for making decisions about testing adaptations in the classrooms. When asked to rate testing adaptations on scales indicating helpfulness to the student and ease of implementation, most of the adaptations rated as most helpful were not rated as easy to make. Examples of adaptations rated most helpful included, "giving individual help with directions on a test" and "simplifying wording of test questions." "Allowing answers in outline formats" and "giving take-home tests" were rated as some of the least helpful adaptations. Items such as "using black-and-white copies instead of dittos" and "giving individual help with directions on a test" were rated as easy adaptations to make while "teaching students test-taking skills" and "allowing word processors" were rated among the most difficult to implement. The majority of teachers (67%) believe that it is unfair to provide testing adaptations only for students with identified disabilities. Many of them stated that adaptations should be made for all students who need them. A small percentage of teachers (8%) believed that adaptations were unfair because they believed that all students in general education classes must work at general education standards.

Two other regional studies of general educators' perceptions were conducted by Gajria, Salend, and Hemrick (1994) and Schumm and Vaughn (1991). Gajria et al. (1994) surveyed sixty-four teachers (grades 7-12) from two suburban school districts in New York using a questionnaire on awareness, use, integrity, effectiveness, and ease of use for 32 test design modifications. Schumm and Vaughn (1991) surveyed 25 elementary, 23 middle school, and 45 high school teachers from one metropolitan school district in the southeastern United States. Teachers in this study rated both the desirability and feasibility of 30 classroom adaptations on a seven-point Likert-type scale using the Adaptation Evaluation Instrument (AEI). Created by the authors, the AEI was designed to investigate teachers' attitudes about the desirability and feasibility of making adaptations for mainstreamed students.

Consistent with the findings of Jayanthi et al. (1996), results from both studies indicated that modifications and adaptations that require little individualization were rated as most feasible and

were most likely to be used by teachers. Conversely, those modifications and adaptations that required changes in planning, curriculum use, or evaluation procedures were rated as least feasible in Schumm and Vaughn's study. In Gajria et al.'s study, modifications involving changes in administrative procedures were less likely to be used than those pertaining to changes in test design. Schumm and Vaughn found that ratings of desirability were significantly higher than ratings of feasibility for all 30 adaptations. Similarly, Gajria et al. found that for one-third of the adaptations, perceived effectiveness was rated significantly higher than use. Schumm and Vaughn reported that the adaptations rated most desirable by teachers were those that related to students' social and motivational adjustment and did not require any curricular or environmental adaptations by the teacher (p. 22). Teachers in this study rated adaptations to materials or instruction as neither desirable nor feasible. Likewise, teachers in the Gajria et al. study were most likely to use modifications that could be applied to all students (e.g., ample spaces for students' responses on the test protocol). They were less likely to use modifications that were specific to the needs of individual students (e.g., adjust reading level of test to meet students' needs).

Although different measures were used in each study to assess teachers' perceptions, results highlight several important themes. In all three studies, teachers were less likely to use, rate as feasible, or rate as easy to implement, modifications or adaptations that required individualization or changes to instruction. According to Schumm and Vaughn, the bottom line for successful inclusion is "teacher willingness to accept and make decisions for students with special needs" (p. 18). The results of their study suggest that this may not be realistic in terms of classroom adaptations. With respect to testing accommodations, the results of Jayanthi et al. (1996) and Gajria et al. (1994) indicate that teachers are familiar with testing modification options, that they are generally responsible for making decisions about accommodations, and that decisions about utilization are linked to perceptions of effectiveness and the resources required for implementation. These findings imply that teachers may not have the knowledge or skills to make individualized adaptations, they may lack information regarding efficient ways to incorporate testing modifications for mainstreamed students, and they may embrace the belief that individualized adaptations should not be made for students. All the authors pointed to the need for teacher education, training, and support if adaptations, and modifications — perceived by teachers to be time-consuming or resource-intensive — are to be truly incorporated into general education classrooms.

The study focusing on student perspectives provides an interesting contrast to these findings. Vaughn, Schumm, Niarhos, and Daugherty (1993) asked 876 middle and high school students about their perceptions of adaptations made by teachers. Results indicated that students preferred teachers who made adaptations, but they had strong feelings about preferred types of adaptations. With respect to instructional practices, most students preferred teachers who were

attentive to individual needs, sensitive to diverse learning patterns, and who adjusted instruction to meet the ability level of the student. Students preferred no adaptations in terms of textbooks, materials, homework, or tests. These results held when students were divided into high- and low-achieving groups. The authors speculated that preferences are related to the appearance of differential treatment. That is, students are less supportive of adaptations that "overtly indicate differential treatment" (p. 115); this effect appears to intensify as students move from middle to high school.

The researchers also surveyed students about achievement and social alienation in order to measure the relationship between these variables and students' perceptions of teacher adaptations. Results of these analyses showed that students who felt more alienated from their peers and teachers were more likely to have favorable views of teachers who make adaptations. Unfortunately, the authors did not indicate whether their sample consisted of students in regular education, special education, or some combination of both; thus, caution must be exercised in reaching generalizations about these findings.

The results of research focusing on teacher and student perceptions of accommodations and adaptations suggest that there are some points of agreement between students and teachers about adaptations (e.g., students don't want differential treatment and teachers are more likely to use modifications that can be used by all students). However, it appears that the driving forces behind these seemingly congruent ratings are probably quite different. Furthermore, there are important domains where students and teachers are nearly polar opposites with respect to their perceptions: whereas students are most interested in individualized instructional practices, teachers report these types of modifications and adaptations are less feasible and less likely to be used.

Conceptual Issues Related to Testing and Accommodations

As mentioned in the introduction, this search for empirical studies of testing accommodations uncovered a host of articles addressing the issue of testing accommodations from a variety of contexts and perspectives, representing such diverse domains as employment testing, WISC-III assessment, performance assessment, and bar examinations. The value of this diverse literature is that it highlights the fact that despite differing contexts, the primary conceptual issues surrounding testing and the provision of accommodations or modifications are generally the same. While there are unique considerations to be made for each type of assessment, there are also themes that continually resurface. In this final section, we review seven of these overarching themes, many of which have been touched on in earlier sections of the report. A summary of the themes and the articles addressing them is provided in Table 3.

Table 3: Major Themes in Literature on Accommodations and Articles that Address Them

Theme	Articles
Validity	Bennett, 1995 Camara & Brown, 1995 Fischer, 1994 Geisinger, 1994 Hishinuma, 1995 Linn, 1994a Linn, 1994b Messick, 1994 Willingham, 1989
Compliance with ADA Legislation	Bennett, 1995 Fischer, 1994 Willingham, 1989
Balance	Hishinuma, 1995 Linn, 1994a Phillips, 1993, 1996 Ragosta, 1991
When and At What Level to Apply Measurement Standards	Camara & Brown, 1995 Linn, 1994a
Eligibility for Accommodations	Fischer, 1994 Hishinuma, 1995 Overton, 1991 Ragosta, 1991
Generalizability of Accommodations	Bennett, 1995 Fischer, 1994 Hishinuma, 1995
Value and Limitations of Tests	Camara & Brown, 1995 Linn, 1994a Linn, 1994b

Validity. When representative literature from other domains was sampled, the issue of validity emerged as a predominant theme; in fact, the concept of validity was discussed in 100% (12/12) of the articles reviewed for this section. Across contexts it appears that consensus exists regarding the centrality of construct validity and of validity as a unified concept (e.g., Camara & Brown, 1995; Linn, 1994a; Messick, 1994). According to Messick (1994), there are six aspects to construct validity, all of which apply to educational and psychological measurement: content, substantive, structural, generalizability, external, and consequential. Under a unified concept of validity, test validity does not rely on nor require any one form of evidence. What is required, according to Messick, is "a compelling argument that the available evidence justifies the test interpretation and use...hence, validity becomes a unified concept and the unifying force is the meaningfulness or trustworthy interpretability of the test scores and their action implications, namely, construct validity" (p. 15). This focus on validity was often accompanied by calls for more research. Several authors noted that although there are standards (e.g., the AERA, APA, NCME Standards) that provide guidance about accommodations, much more empirical data are needed (Bennett, 1995; Fischer, 1994; Willingham, 1989). The findings presented earlier in this report lend support to the need for empirical research. Hishinuma (1995) expresses this sentiment well when he states that "legislative intent goes well beyond any preexisting research knowledge of the psychometric effects of accommodations" (p. 134). It was also noted by many of the authors that issues of fairness, opportunity-to-learn, and the social consequences of test use are receiving increased attention with respect to evaluating test validity (e.g., Camara & Brown, 1995; Geisinger, 1994; Linn, 1994b; Messick, 1994).

Compliance with ADA legislation. Most of the 12 documents made mention of the significance of ADA and its impact on testing practices, including implications for flagging scores (e.g., Bennett, 1995; Fischer, 1994; Willingham, 1989). For example, Fischer (1994) devoted an entire article to the discussion of what ADA requires of assessment programs and what the measurement implications of this act are. While discussions surrounding this theme seem to focus on the significant changes ADA mandates in terms of accommodations for testing, Fischer asserts that "the new law requires nothing more than what good professional practice already requires: that each test and assessment device be valid for all examinees" (p. 18).

Balance. This issue was raised earlier by Phillips (1993, 1996). Across disciplines, authors referred to the tensions that exist in trying to comply with ADA and provide reasonable accommodations while at the same time upholding measurement standards that maintain the validity of the assessment measure without providing an unfair advantage for people with disabilities (e.g., Hishinuma, 1995; Linn, 1994a; Ragosta, 1991).

When and at what level to apply measurement standards. In general, there appears

to be consensus that the application of standards is completely dependent upon the purposes and interpretational uses of the test (e.g., Camara & Brown, 1995; Linn, 1994a). Measurement standards should be most stringent when decisions are being made about an individual. In other words, as the stakes associated with an assessment increase, so too should measurement standards. Linn (1994a) notes however, that as important as measurement standards are, they generally take a backseat for policy makers who are more interested in issues of cost and the impact of assessments on various groups of people.

Eligibility for accommodations. There appears to be agreement that not all students with disabilities need or should receive special test accommodations; thus, the question becomes how to make eligibility decisions. According to Ragosta (1991), the need for special accommodations is "related to the severity of the disability and whether the disability would negatively affect a candidate's test score" (p. 12). She suggests that eligibility guidelines for special accommodations of the bar exam for students with learning disabilities could reasonably include consideration of: (1) the timing of diagnosis, (2) past educational practice, and (3) accommodations that are available in the profession. In the employment context, Fischer (1994) suggests that documentation from a health-care professional or other official evidence of a documented diagnosis is appropriate. Related both to eligibility and the need for data is the need articulated by several of the authors for clear guidelines for making decisions about accommodations (e.g., Hishinuma, 1995; Overton, 1991; Ragosta, 1991).

Generalizability of accommodations. The general consensus on this issue seems to be that no single accommodation fits all (e.g., Fischer, 1994), yet several authors spoke of the ability to make some generalizations and possibly develop a continuum of accommodations (e.g., Hishinuma, 1995). There was some disagreement about this issue, however. For example, Bennett (1995), in describing computer-based testing (CBT), asserts that generalized accommodations that benefit everyone are desirable and feasible.

Value and limitations of tests. There were several discussions (e.g., Camara & Brown, 1995; Linn, 1994a) about the evolving uses of tests today. Camara and Brown (1995) state that tests generally serve three purposes today, both in education and employment: (1) decision making, (2) aiding instruction, and (3) accountability (Resnick & Resnick, 1991, cited in Camara & Brown, 1995). They contend that policy makers may envision assessment devices as "the principle means of jump-starting educational reform by attaching high stakes in a deliberate effort to drive curriculum" (p. 9), but assert that "using tests as agents of change represents a fundamental change in the purpose of measurement and assessment, as well as a somewhat inflated notion of what tests are and what they can do" (p. 10). Additional research suggests that although many policy makers are anxious to use assessments as a policy tool, they are less sure about the specific uses for the resulting data (Linn, 1994b). This is particularly troubling given that uses and inferences are supposed to drive the measurement process.

Summary

The purpose of this report was to provide an updated review of the literature on testing accommodations for students with disabilities, with a particular emphasis on studies examining the effects of testing accommodations on the technical integrity of assessment measures. Like the original Thurlow, Ysseldyke, and Silverstein (1993) report, we sought to answer the question, "What do we currently know about testing accommodations for students with disabilities?" In this review we found only six studies published since the 1993 report that examined the effects of various accommodations or modifications. Overwhelmingly, the results of these studies — varying widely in purpose, design, and level of rigor — pointed to the need for more research, both in terms of quantity and quality.

Without a comprehensive empirical base from which to draw, this review also encompassed other pertinent literature bases, including legal issues related to testing accommodations, teacher and student perceptions of accommodations, as well as conceptual issues pertaining to assessment and accommodations or modifications. These literature bases are important for several reasons. They highlight the fact that despite an absence of empirical studies, there is considerable effort being directed toward this topic in the form of theory, practice, and research (with the establishment of the projects such as those listed in Appendix B). Both the literature on legal issues and the research on teacher and student perceptions provide some guidance for practitioners and policy makers who are currently wrestling with how best to accommodate students with disabilities in testing situations. The fact that consistent themes emerge across disciplines with respect to testing accommodations suggests that collaborative research efforts are possible and are a potential area to explore for future investigations.

To the question, "What do we know about testing accommodations for students with disabilities?" the answer seems to be that we have gained only a little more empirical information. Nevertheless, we have a much richer understanding of the relevant issues and are now poised to develop methodologically sound, empirically driven research studies that will be useful for practitioners and policy makers and, ultimately, be beneficial for students with and without disabilities.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Americans with Disabilities Act (A.D.A., 1990). Pub. L. No. 101-336, 42 U.S.C. 12101 *et seq*.

Bennett, R.E. (1995). *Computer based testing for examinees with disabilities: On the road to generalized accommodation* (RM-95-1). Princeton, NJ: Educational Testing Service.

Camara, W.J. & Brown, D.C. (1995). Educational and employment testing: Changing concepts in measurement and policy. *Educational Measurement, 14* (1), 5-11.

Dalton, B., Morocco, C.C., Tivnan, T., & Rawson, P. (1994). Effect of format on learning disabled and non-learning disabled students' performance on a hands-on science assessment. *International Journal of Educational Research, 21* (3), 229-316.

Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979), *aff'd in part, rev'd in part*, 644 F.2d 397 (5th Cir. 1981); *on remand*, 564 F. Supp. 177 (M.D. Fla. 1983), *aff'd*, 730 F.2d 1405 (11th Cir. 1984).

Dunlap, G., Foster-Johnson, L., Clarke, S., Kern, L., & Childs, K.E. (1995). Modifying activities to produce functional outcomes: Effects on the problem behaviors of students with disabilities. *Journal of the Association for Persons with Severe Handicaps, 20* (4), 248-258.

Erickson, R., Thurlow, M., & Ysseldyke, J. (1996). *Neglected numerators, drifting denominators, and fractured fractions: Issues in determining participation rates for students with disabilities in statewide assessment programs*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Fischer, R.J. (1994). The Americans with Disabilities Act: Implications for measurement. *Educational Measurement: Issues and Practices, 13* (3), 17-26.

Gajria, M., Salend, S.J., & Hemrick, M.A. (1994). Teacher acceptability of testing modifications for mainstreamed students. *Learning Disabilities Research and Practice, 9* (4), 236-243.

Geisinger, K.F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education, 7* (2), 121-140.

Hawaii State Dept. of Educ., 17 EHLR 360 (O.C.R. 1990).

Hishinuma E.S. (1995). WISC-III accommodations: The need for practitioner guidelines. *Journal of Learning Disabilities, 28* (3), 130-135.

Jayanthi, M., Bursuck, W.D., Havekost, D.M., Epstein, M.H., & Polloway, E.A. (1994). School district testing policies and students with disabilities: A national survey. *School Psychology Review, 23* (4), 694-703.

Jayanthi, M., Epstein, M.H., Polloway, E.A., & Bursuck, W.D. (1996). A national survey of general education teachers' perceptions of testing adaptations. *The Journal of Special Education, 30* (1), 99-115.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education*, Washington, DC: Author.

Linn, R.L. (1994a). Evaluating the technical quality of proposed national examination systems. *American Journal of Education, 102* (4), 565-580.

Linn, R.L. (1994b). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23* (9), 4-14.

Maxwell v. Pasadena I.S.D., No. 92-017184, 295th District Court of Harris County, TX, Dec. 29, 1994.

Messick, S. (1994). *Alternative modes of assessment, uniform standards of validity*. Princeton, N.J.: Educational Testing Service (ED 380 504).

Mick, L.B. (1989). Measurement effects of modifications in minimum competency test formats for exceptional students. *Measurement and Evaluation in Counseling and Development, 22*, 31-36.

Munger, G.F. & Loyd, B.H. (1991). Effect of speededness on test performance of handicapped and nonhandicapped examinees. *Journal of Educational Research, 85* (1), 53-57.

Overton, G.R. (1991, February). Accommodation of disabled persons. *The Bar Examiner*, 6-10.

Perlman, C., Borger, J., Collins, C., Elenbogen, J., & Wood, J. (1996, April). *The effect of extended time limits on learning disabled students' scores on standardized reading tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, New York.

Phillips, S.E. (1993). Testing accommodations for disabled students. *Education Law Reporter, 80*, 9-32.

Phillips, S.E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7* (2), 93-120.

Phillips, S.E. (1995). *All students, same test, same standards: What the new Title I legislation will mean for the educational assessment of special education students*. Oak Brook, IL: North Central Regional Educational Laboratory.

Phillips, S.E. (1996). Legal defensibility of standards: Issues and policy perspectives. *Educational Measurement: Issues and Practice, 15* (2), 5-19.

Pomplun, M. (1996). Cooperative groups: Alternative assessment for students with disabilities. *The Journal of Special Education, 30* (1), 1-17.

Ragosta, M. (1980). *Handicapped students and the SAT* (Research Report 80-12). Princeton, NJ: Educational Testing Service.

Ragosta, M. (1991). Testing bar applicants with learning disabilities. *The Bar Examiner, February*, 11-15.

Ragosta, M. & Wendler, C. (1992). *Eligibility issues and comparable time limits for disabled and nondisabled SAT examinees*. New York, N.Y.: College Entrance Examination Board Report No. 92-5. (ED 349 337).

Schumm, J.S. & Vaughn, S. (1991). Making adaptations for mainstreamed students: General classroom teachers' perspectives. *Remedial and Special Education, 12* (4), 18-27.

Salend, S.J. (1995). Modifying tests for diverse learners. *Interventions in Schools and Clinics, 31* (2), 84-90.

Section 504 of the Rehabilitation Act, 29 U.S.C. 701 *et seq*. (1973).

Thurlow, M.L., Ysseldyke, J.E., & Silverstein, B. (1993). *Testing accommodations for students with disabilities: A review of the literature* (Synthesis Report 4). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Vaughn, S., Schumm, J.S., Niarhos, F.J., & Daugherty, T. (1993). What do students think when teachers make adaptations? *Teaching and Teacher Education, 9* (1), 107-118.

Wards Grove Packing Co. v. Antonio, 109 S.Ct. 2115 (1989).

Willingham, W.W. (1989). Standard testing conditions and standard score meaning for handicapped examinees. *Applied Measurement in Education, 2* (2), 97-103.

Appendix A

A Review of Early Studies in Testing Accommodations Conducted by the American College Testing Program (ACT) and the Educational Testing Service (ETS)

A Review of Early Studies in Testing Accommodations Conducted by the American College Testing Program (ACT) and the Educational Testing Service (ETS)

In 1984, a report on issues pertaining to participation in the ACT assessment by examinees with disabilities was produced by the American College Testing Program (Laing & Farmer, 1984). The report summarized some information gathered from ACT's records from 1978-79 through 1982-83. Five groups of examinees were considered: students without disabilities and students with disabilities who took the exam in a standard administration, and students with visual impairments, hearing impairments, or motor disabilities (identified as including physical and learning disabilities) who took a nonstandard administration.

In order to be permitted to take the ACT assessment under nonstandard conditions, persons with disabilities must be professionally diagnosed, and proper documentation of the disability must be sent to ACT. Diagnosis and certification of the disability must be provided by a qualified professional with appropriate credentials.

Among the accommodations ACT offers are: extended time, large type, Braille, audio cassette editions of the test, the use of a reader, assistance in filling out the answer folder, and the signing of instructions. Furthermore, individuals with disabilities are allowed to bring to the exam selected assistive devices such as a Braille, slate and stylus, magnifying glass, or tape recorder.

Predictive validity was examined using first-year college grades as the criterion measure. It was reported that the prediction of first-year college GPA was about equally accurate for examinees without disabilities and examinees with disabilities, when both groups took the exam under standard testing conditions. For both, the correlation between predicted and actual first year college GPA was .59 (Maxy & Levitz, 1980 in Laing & Farmer, 1984). For examinees with visual disabilities who were tested under nonstandard conditions, the correlation between predicted and earned grades was .52; for students with motor (physical and learning) disabilities, the correlation was .39. The sample of students with auditory disabilities was too small (n = 9) to draw conclusions. It should be noted that the regression equations in all of the above cases were established on data from regularly tested examinees.

The American College Testing (ACT) patterns resemble those found through other efforts conducted by the Educational Testing Service (ETS). ETS conducted a series of studies on the comparability of standard and nonstandard versions of the Scholastic Aptitude Test (SAT) and the Graduate Record Examinations (GRE) General Test. Findings are reported in an entire book on the topic of testing people with disabilities (Willingham, Ragosta, Bennett, Braun, Rock, & Powers, 1988). In these studies, researchers focused on test comparability for four

groups of students with disabilities: those with hearing impairments, learning disabilities, physical disabilities, and visual impairments.

In general, test comparability is analyzed to determine whether tests are fair for different subgroups, such as various ethnic groups. Modified tests or testing conditions deviate from standardization to some degree in order to remove sources of irrelevant difficulty. Consequently, Willingham et al. (1988) argued that comparability in these cases must be broken down into score comparability and task comparability.

Willingham et al. (1988) define both of these terms. Score comparability referred to "comparable meaning and interpretation of test performance, not necessarily the same distribution of scores for different groups" (p. 13). Willingham et al. identified five respects in which scores should be generally comparable: reliability, factor structure, item functioning, predicted performance, and admission decisions. Task comparability was used to mean that there are equivalent cognitive demands made on different groups (e.g., those with disabilities and those without disabilities), not necessarily that the superficial characteristics of the test situation are the same. Critical questions to consider are: Is the content comparable? Is the timing for examinees with disabilities comparable to that for examinees without disabilities? (Willingham et al., 1988)

How can both score comparability and task comparability be evaluated? Score comparability can be evaluated empirically. Task comparability, on the other hand, is evaluated primarily through judgments of people with disabilities and professionals who work with them. In the ETS studies, eight specific indicators of comparability (five score comparability and three task comparability indicators) were studied:

Score Comparability

- Reliability
- Factor structure
- Differential item functioning
- Prediction of performance
- Admissions decisions

Task Comparability

- Test content
- Testing accommodations
- Test timing

Findings on each of these indicators are detailed in the following paragraphs.

Reliability

ETS researchers found that nonstandard and standard versions of both the SAT and GRE had equivalent reliability (Bennett, Rock, & Jirele, 1986, 1987; Bennett, Rock, & Kaplan, 1985, 1988; Bennett, Rock, Kaplan, & Jirele, 1988). The nonstandard version that they evaluated included Braille, cassette recorded, and large type editions of the tests. There was some evidence that different sections of the SAT were not as highly correlated for students with disabilities as for students without disabilities (e.g., quantitative and verbal abilities sections), but in general similar correlations were found among sections for students with and without disabilities.

Factor structure

Factor structures of the standard and nonstandard examinations for the SAT were quite similar, thus supporting the assumption that the cognitive abilities assessed by nonstandard tests are comparable to those assessed by standard measures (Rock, Bennett, & Kaplan, 1987). For the GRE, a four-factor model fit better than a three-factor model. The three-factor model had particular problems in fit for students with visual impairments who were taking a large-type test and for examinees with physical disabilities who were taking a standard test administration. Specifically, the item types that made up the analytical factor did not appear to function effectively as a single factor. The researchers concluded that these results suggest that analytical scores and total scores might have different meaning for groups with and without disabilities (Rock, Bennett, & Jirele, 1988).

Differential item functioning

In general, test item difficulty was similar for individuals with and without disabilities on both the SAT and the GRE. The one exception to this appeared on the Braille version of mathematical portion of the SAT, where a few items were more difficult for the examinees taking the Braille version of the test (Bennett, Rock, & Kaplan, 1985, 1987).

Prediction of performance

One area where test comparability appeared to be questionable was the prediction of academic performance. When nonstandard test scores were used alone, they tended to be less valid predictors of academic performance than were standard test scores for examinees without disabilities. Further, the predictability of academic performance of different subgroups of students with disabilities varied. Test scores substantially underpredicted college grades for students with hearing impairments who had enrolled in colleges that provided them with special services. In contrast, SAT scores overpredicted college performance for students with physical

handicaps and learning disabilities (Braun, Ragosta, & Kaplan, 1986). It should be noted that when supplemented with grade point averages, nonstandard tests did not consistently over or underpredict academic performance for students with disabilities as a whole. Students with disabilities who had low test scores and prior grades, however, tended to do somewhat better in college than predicted, while those with high scores on both tended to do somewhat worse than predicted.

Admissions decisions

Overall, admissions decisions for students with disabilities were comparable to decisions for students without disabilities. The effect of flagging (i.e., identifying test scores from nonstandard administrations) seemed minimal (Benderson, 1988). However, there were three subgroups of applicants with disabilities whose actual rate of admissions differed significantly from what was predicted for them. Applicants with hearing impairments were significantly more likely to be admitted; students with learning disabilities who ranked in the mid- to upper-range among applicants at the college to which they applied were slightly less likely to be admitted; and, for a relatively small number of applicants with visual and physical disabilities who were applying to smaller institutions, the admissions were lower than predicted. ETS researchers hypothesized that this finding was a consequence of the higher probability that smaller institutions are less able to provide the needed resources or special equipment for individuals with visual and physical impairments (Willingham et al., 1988).

Test content

The issue of test content is related to concerns about whether students with disabilities and students without disabilities take essentially the same test. In other words, does the student's disability place different task demands on the test? Willingham et al. (1988) identified three types of information that aid in determining task comparability: (1) analyzing items and factors in the test through statistical methodology, (2) the opinions of students with disabilities who took the nonstandard test, and (3) the relative performance on different test sections.

Students with disabilities scored relatively higher on the verbal than on the mathematical sections of the SAT and GRE despite the fact that many of those students with disabilities reported having greater difficulty with the vocabulary and amount of reading material on the test compared to the mathematical sections (as did many of the other students). This included students with learning disabilities, for whom one would expect relatively greater difficulty with reading (Willingham et al., 1988), but not students with hearing impairments.

Willingham et al. (1988) concluded that while the task demands of the admissions test are more difficult for some students with disabilities than for students without disabilities, the test

content overall appears to be comparable. He makes two suggestions: (1) look into the feasibility of a manual translation of the test for students who are deaf, and (2) try to eliminate the mathematical items that are differentially difficult for students who take a Braille version of the test.

Testing accommodations

Among the test accommodations ETS offers are alternative test formats (e.g., Braille, cassette, large type), alternative ways to record answers, separate test locations, and extra time (ETS, 1990).

Test timing

Evidence of noncomparability of task in the standard and nonstandard versions of the SAT and GRE was found in the test indicator. Willingham et al. (1988) stated that examinees with disabilities were more likely to finish the test than examinees without disabilities. They also reported that some test items near the end of the examinations were relatively easier for some groups of students with disabilities than for others. Related to this was the finding that in some instances college performance was overpredicted by test scores based on considerably extended testing time. Extended time for students with learning disabilities was identified as a particularly difficult issue. Allowing these students extra time is controversial because students are defined as having a learning disability when they exhibit low academic performance in school and lower performance on achievement tests than on ability tests.

Recommendations

On the basis of its research on special administrations of the SAT and GRE, ETS made several recommendations. The recommendations primarily address the use of test scores obtained from nonstandard administrations, not the issue of whether or which accommodations are appropriate. Based on the findings of its researchers, ETS suggested that users of nonstandard scores should:

- Use multiple criteria to predict academic performance of disabled students.
- Give less weight to traditional predictors and more consideration to the student's background and nonscholastic achievements.
- Avoid score composites.

- Avoid the erroneous belief the nonstandard scores are systematically either inflated or deflated.

- Where feasible and appropriate, report scores in the same manner as those obtained from standard administrations (ETS, 1990, Executive Summary).

These recommendations were based on findings similar to those found for the ACT (Laing & Farmer, 1984). In both the ETS and the ACT research, nonstandard testing of students with disabilities resulted in lower correlations between test scores and first-year college GPA. Similarly, both tests tended to overpredict grades for students with physical handicaps and learning disabilities.

In 1991, ETS initiated an effort to examine the possibilities and problems of another testing accommodation — the use of computer-based testing. The possibilities for adaptations are wide ranging when computer technology is explored, including, for example, videodisc systems that display written text simultaneously with an inset of a person translating the text into sign language, voice synthesizers that simulate speech for individuals who are blind, and movement controls that allow a person with difficulty speaking and limited hand movement to both enter text and respond to text presented on the monitor. ETS found that the challenge of testing goes beyond the mere taking of the test: "every aspect of the testing process, from registration to score reporting, may present impediments to people with disabilities" (ETS, 1992, p. 7).

Through the use of computer-based testing, researchers at ETS see the possibility of addressing many of the issues facing testing programs. They suggest that computer-based tests can be designed "from the outset in ways that do not present barriers for individuals with disabilities" (p. 7). In line with this view, ETS introduced a computerized GRE in October 1992, and has started working on a computerized version of the SAT. Despite these advances, many questions still exist about the use of computerized testing in general. For example, the National Center for Fair & Open Testing recently produced a "fact sheet" that highlights some of the questions surrounding computerized testing (Fair Test, 1993). Noting that "the new tests are being ushered in before adequate evidence of either their comparability to current exams or their fairness have been collected," Fair Test highlights the following as just some of the unresolved problems of computerized testing:

- Inadequate support exists for claims that scores of computerized and pencil-and-paper tests are equivalent.

- Computerized tests constrain users because they cannot underline, scratch out eliminated choices, or scan materials in the same way they can with paper and pencil tests.

- Computer screens may take longer to read, and it may be more difficult to detect errors on computer screens.

References

Benderson, A. (Ed.) (1988) Testing, equality, and handicapped people. *Focus* (ETS Publication; Princeton, NJ), *21* (23 pp).

Bennett, R. E., Rock, D. H., & Jirele, T. (1987). GRE score level, test completion, and reliability for visually impaired, physically handicapped, and nonhandicapped groups. *The Journal of Special Education, 21* (3), 9-21.

Bennett, R. E., Rock, D. A., & Jirele, T. (1986, February). *The psychometric characteristics of the GRE General Test for three handicapped groups* (ETS Research Report RR-86-6). Princeton, NJ: Educational Testing Service.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1985, November). *The psychometric characteristics of the SAT for nine handicapped groups* (ETS Research Report RR-85-49). Princeton, NJ: Educational Testing Service.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement, 24* (1), 44-55.

Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1988). Level reliability and speededness of SAT scores for nine handicapped groups. *Special Services in the Schools, 4* (3/4), 37-54.

Bennett, R. E., Rock, D. A., Kaplan, B. A., & Jirele (1988). Psychometric characteristics. In W. W. Willingham, M. Ragosta, R. E. Bennet, H. Braun, D. A. Rock, & D. E. Powers (Eds.) *Testing handicapped people* (pp. 83-97). Boston: Allyn & Bacon.

Braun, H., Ragosta, M., & Kaplan, B. (1986, November). *The predictive validity of the GRE General Test for disabled students* (ETS Research Report 86-42). Princeton, NJ: Educational Testing Service.

Braun, H., Ragosta, M., & Kaplan, B. (1986, October). *The predictive validity of the scholastic aptitude test for disabled students* (ETS Research Report RR-86-38). Princeton, NJ: College Entrance Examination Board, Educational Testing Service, Graduate Record Examinations Board.

ETS. (1990). Testing persons with disabilities: A report for ETS programs and their constituents. Princeton, NJ: Educational Testing Service.

ETS. (1992). ETS conference examines the technology of computer-based testing for people with disabilities. *ETS Developments*.

FairTest. (1993). *Computerized testing: More questions than answers*. Cambridge, MA: National Center for Fair & Open Testing.

Laing, J., & Farmer, M. (1984). *Use of the ACT assessment by examinees with disabilities* (Research Report No. 84). Iowa City, IA: American College Testing Program (ACT).

Maxey, E. J., & Levitz, R. S. (1980, April). *ACT services for the handicapped*. Paper presented at the meeting of the American Association of Collegiate Registrars and Admissions Officers, New Orleans, LA.

Rock, Bennett, & Jirele (1988). Factor structure of the Graduate Record Examinations general test in handicapped and nonhandicapped groups. *Journal of Applied Psychology, 73* (3), 383-392.

Rock, Bennett, & Kaplan (1987). Internal construct validity of a college admissions test across handicapped and nonhandicapped groups. *Educational and Psychological Measurement, 47*, 193-205.

Willingham, W. W. (1988). Introduction. In W. W. Willingham, M. Ragosta, R. E. Bennett, H. Braun, D. A. Rock, & D. E. Powers (Eds.), *Testing handicapped people*. Boston: Allyn & Bacon.

Willingham, W. W., Ragosta, M, Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (Eds.). (1988). *Testing handicapped people*. Boston: Allyn & Bacon.

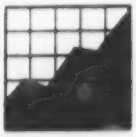
Appendix B

Research Projects Supported by the U.S. Office of Special Education Programs (OSEP) and the U.S. Office of Educational Research and Improvement (OERI)

Research Projects Supported by the U.S. Office of Special Education Programs (OSEP) and the U.S. Office of Educational Research and Improvement (OERI)

OSEP Projects	*Recipient Organizations*
Examining Alternatives for Outcome Assessment for Children with Disabilities	Maryland State Department of Education
Performance Assessment and Standardized Testing for Students with Disabilities: Psychometric Issues, Accommodation Procedures, and Outcome Analyses	Wisconsin Center for Education Research
Project Reading ABC: An Alternative Reading Assessment Battery for Children with Severe Speech and Physical Impairments	Center for Literacy and Disability Studies - University of North Carolina

OERI Projects	*Recipient Organizations*
Inclusive Comprehensive Assessment System	Delaware Department of Public Education
The Maryland Assessment System Project	Maryland State Department of Education
Grade 5 and 8 Integrated Social Studies Statewide Assessment Project	Michigan Department of Education
Minnesota Assessment Project	Minnesota Department of Education
Assessment of Media Literacy	North Carolina Department of Public Instruction
North Dakota Language Arts Assessment	North Dakota Department of Public Instruction
Oregon Assessment Development and Evaluation Project	Oregon Department of Education
Pennsylvania Assessment Through Themes Project	Pennsylvania Department of Education
State Collaborative on Assessment and Student Standards (SCASS) Technical Guidelines for Performance Assessment	Council of Chief State School Officers (CCSSO)



NATIONAL CENTER ON EDUCATIONAL OUTCOMES

University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
612/626-1530 • Fax 612/624-0879
http://www.coled.umn.edu/NCEO

The College of Education & Human Development

UNIVERSITY OF MINNESOTA

BEST COPY AVAILABLE